Chapter_3_homework

Christophe

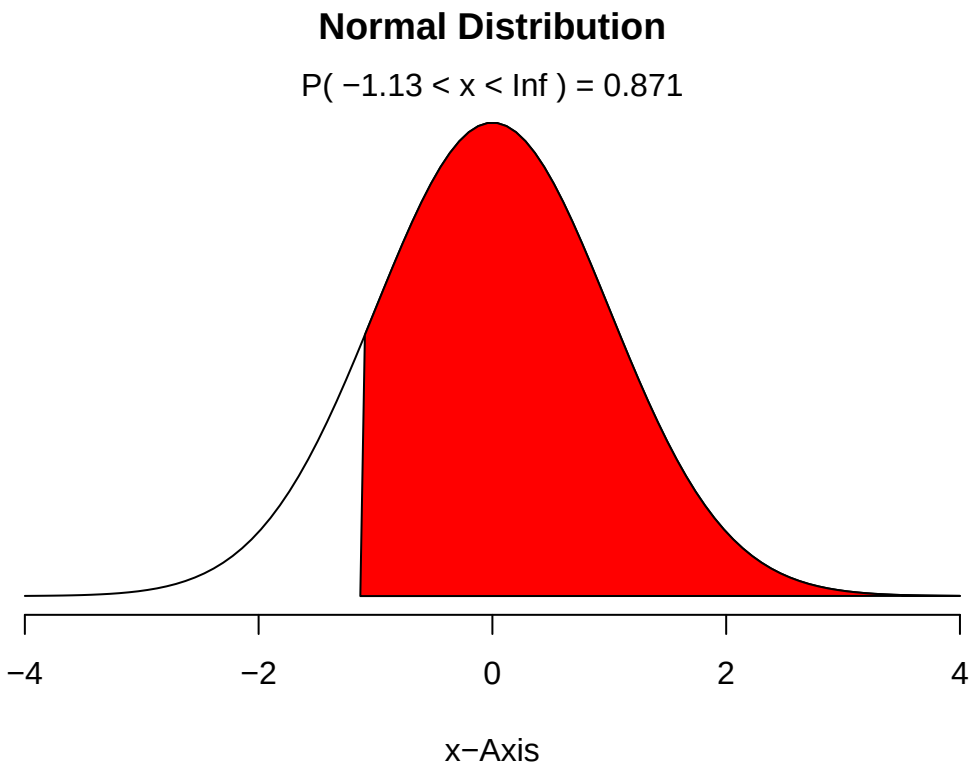
September 17, 2015

3.2 Area under the curve, Part II.

3.2 Area under the curve, Part II. What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z > -1.13$ (b) $Z < 0.18$ (c) $Z > 8$ (d) $|Z| < 0.5$

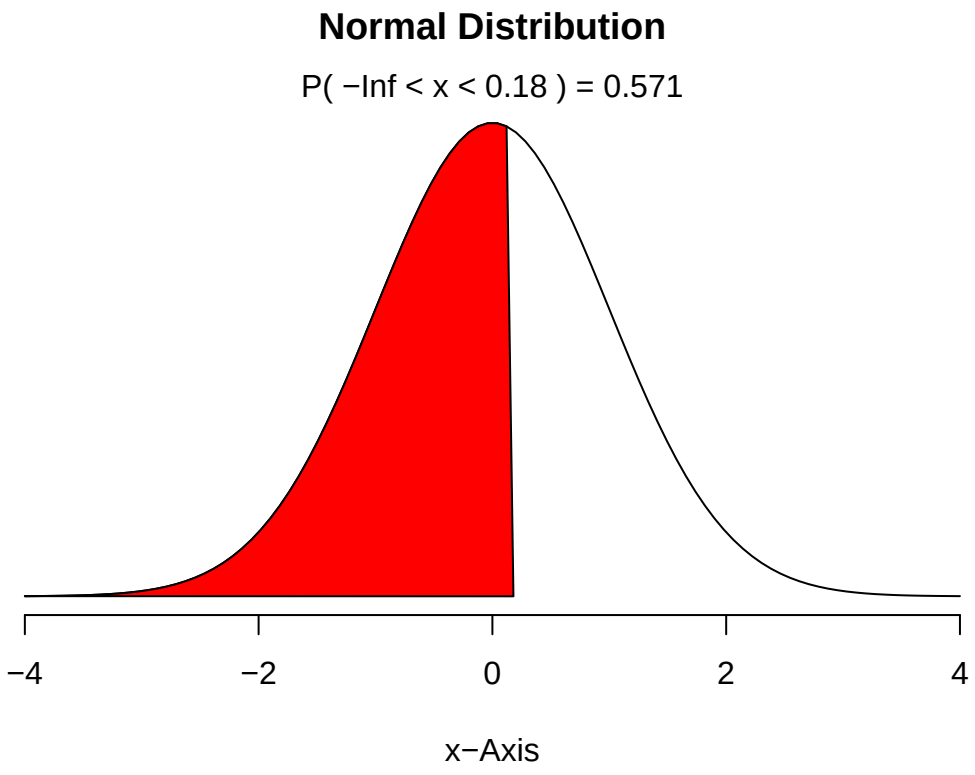
```
library(Rcpp)
library(IS606)

#a
IS606::normalPlot(0, 1, c(-1.13, Inf))
```




```
1 - pnorm(-1.13, mean = 0, sd = 1)
```

[1] 0.8707619

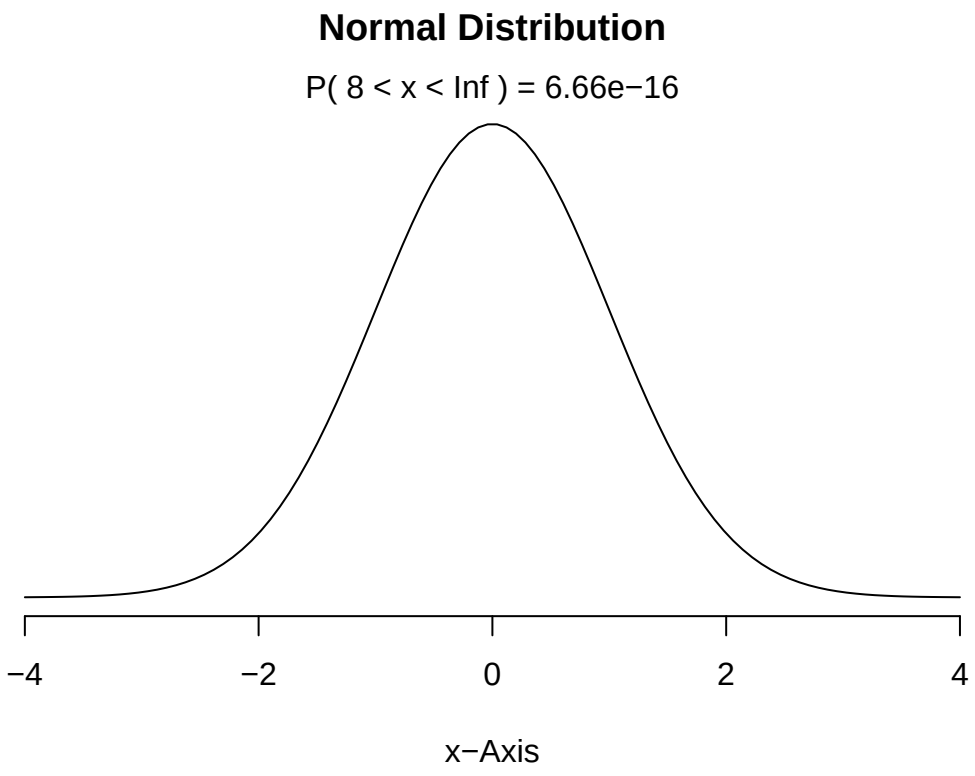
```
#b
IS606::normalPlot(0, 1, c(-Inf, .18))
```




```
pnorm(.18, mean = 0, sd = 1)
```

[1] 0.5714237

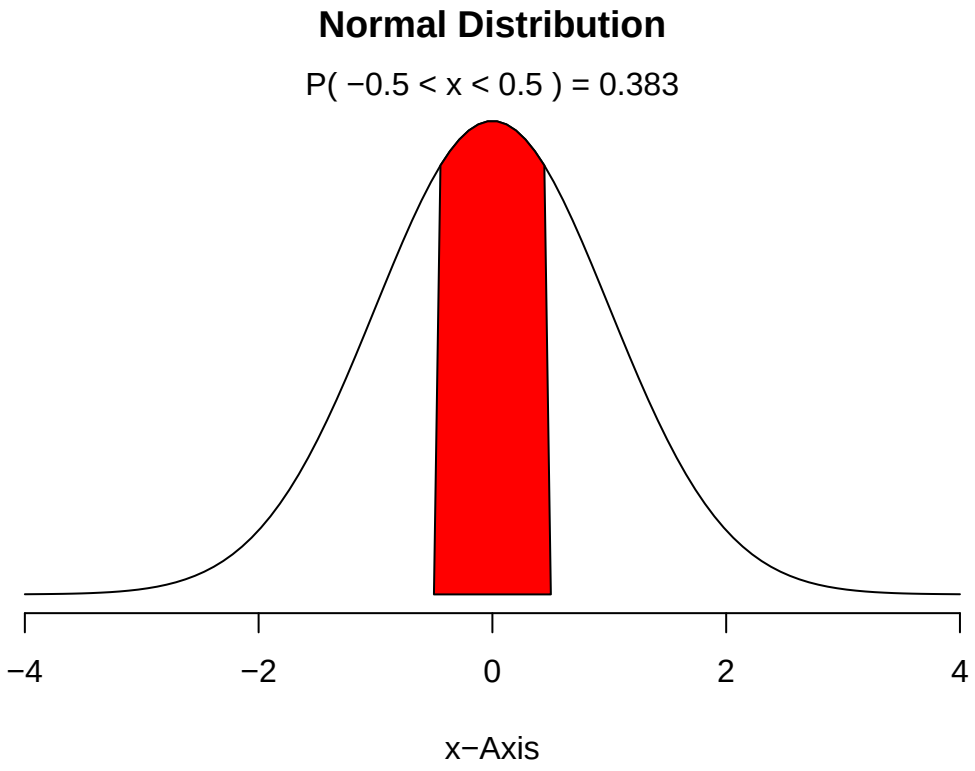
```
#c
IS606::normalPlot(0, 1, c(8, Inf))
```




```
options(scipen=999)
1 - pnorm(8, mean = 0, sd = 1)
```

[1] 0.00000000000000006661338

```
options(scipen=0)

#d
IS606::normalPlot(0, 1, c(-.5, .5))
```




```
pnorm(.5, mean = 0, sd = 1) - pnorm(-.5, mean = 0, sd = 1)
```

[1] 0.3829249

3.4 Triathlon times, Part I.

In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the Men, Ages 30 - 34 group while Mary competed in the Women, Ages 25 - 29 group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

- The finishing times of the Men, Ages 30 - 34 group has a mean of 4313 seconds with a standard deviation of 583 seconds.

- The finishing times of the Women, Ages 25 - 29 group has a mean of 5261 seconds with a standard deviation of 807 seconds.

- The distributions of finishing times for both groups are approximately Normal.

  Remember: a better performance corresponds to a faster finish.

(a) Write down the short-hand for these two normal distributions.

$$MenDistribution = N(\mu = 4313, \sigma = 583)$$
$$WomenDistribution = N(\mu = 5261, \sigma = 807)$$

(b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?

```
Leo_time <- 4948
men_mean <- 4313
men_sd <- 583
Leo_z <- (Leo_time - men_mean) / men_sd
paste("Leo's Z Score is", round(Leo_z,3))
```

[1] "Leo's Z Score is 1.089"

```
Mary_time <- 5513
women_mean <- 5261
women_sd <- 807
Mary_z <- (Mary_time - women_mean)/ women_sd
paste("Mary's Z score is", round(Mary_z,3))
```

[1] "Mary's Z score is 0.312"

> The z scores tell us that Leo's time is actully much worse according to the distribution than Mary's time. This is useful because Leo's time appears much better than Mary's time but when you account for the population distribution its clear that Leo's is not as great as it seems.

(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

> Mary would rank higher in her respective group than Leo would, the reason being that Mary is closer to the mean for her respective group than Leo is. He has a better time in general but when accouting for the distribution within their groups Mary outperforms Leo.

(d) What percent of the triathletes did Leo finish faster than in his group?

```
library(scales)
Leo_better_than <- (1 - pnorm(Leo_time, mean = men_mean, sd = men_sd))
paste("Leo performed better than", percent(Leo_better_than),
      "of those within his group")
```

[1] "Leo performed better than 13.8% of those within his group"

(e) What percent of the triathletes did Mary finish faster than in her group?

```
Mary_better_than <- (1 - pnorm(Mary_time, mean = women_mean, sd = men_sd))
paste("Mary performed better than", percent(Mary_better_than),
      "of those within her group")
```

[1] "Mary performed better than 33.3% of those within her group"

(f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

> The z scores wouldn't change because we can calculate Z scores for non-normal distributions, so (b) would remain the same. However, my answer to (c), (d), and (e) would change because I would not be able to use the normal probability methods for a non-normal distribution. The methods assume a normal distribution so a skewed distribution would cause the normal distribution methods to fail in accuracy.

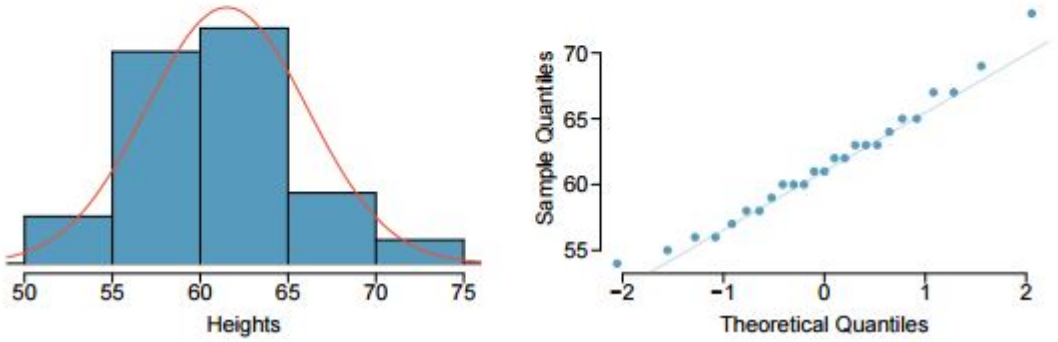
3.18 Heights of female college students

3.18 Heights of female college students. Below are heights of 25 female college students.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
54,	55,	56,	56,	57,	58,	58,	59,	60,	60,	60,	61,	61,	62,	62,	63,	63,	63,	64,	65,	65,	67,	67,	69,	73

(a) The mean height is 61.52 inches with a standard deviation of 4.58 inches. Use this information to determine if the heights approximately follow the 68-95-99.7% Rule.

(b) Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.




```
heights <- c(54,55,56,56,57,58,58,59,60,60,60,61,61,62,62,63,53,53,64,65,65,67,67,69,76)

one_deviation <- subset(heights,
                heights < (mean(heights) + sd(heights)) &
                heights > (mean(heights) - sd(heights)))

paste("The percent of values within one deviation is",
      percent(length(one_deviation)/length(heights)))
```

[1] "The percent of values within one deviation is 68%"

```
two_deviation  <- subset(heights,
                  heights < (mean(heights) + (2* sd(heights))) &
                  heights > (mean(heights) - (2 * sd(heights))))

paste("The percent of values within two deviations is",
      percent(length(two_deviation)/length(heights)))
```

[1] "The percent of values within two deviations is 96%"

```
three_deviation <- subset(heights,
                   heights < (mean(heights) + (3 * sd(heights))) &
                   heights > (mean(heights) - (3 * sd(heights))))

paste("The percent of values within three deviations is",
      percent(length(three_deviation)/length(heights)))
```

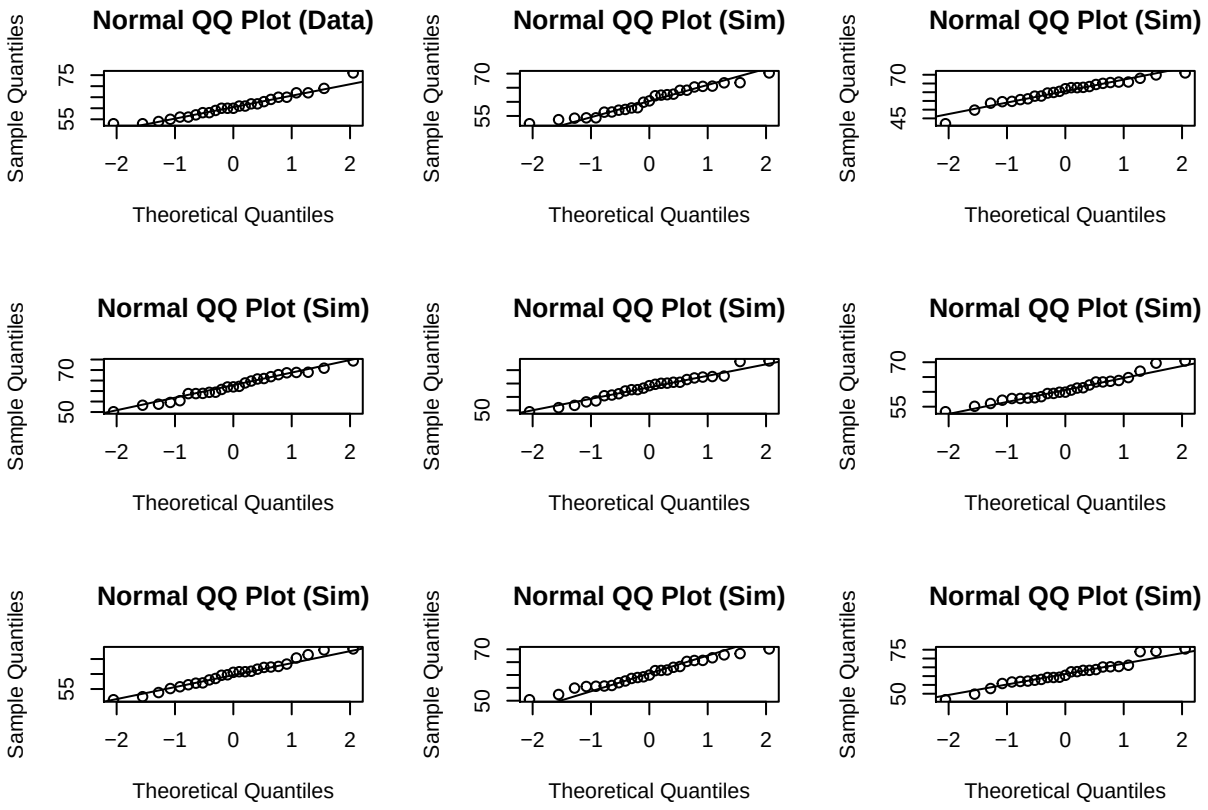
[1] "The percent of values within three deviations is 100%"

(a) The mean height is 61.52 inches with a standard deviation of 4.58 inches. Use this information to determine if the heights approximately follow the 68-95-99.7% Rule.

> The values are very close to the 68-95-99.7% rule, as it currently is 68%-96%-100% so we can conclude it does follow that rule.

(b) Does the data appear to follow a normal distribution? Explain your reasoning using the graphs provided above.

```
qqnormsim(heights)
```



The random qqplot does provide information that the values fall on the line, representing a near normal distribution. The distribution is unimodal and symmetical, it does exhibit some positive skewness but it is very slight. The normal curve does follow the distribution smoothly. The points in the normal probability plot are aligned with the line, with only a few exceptions for outliers. We can therefore conclude that the distribution is normal.

3.22 Defective rate.

A machine that produces a special type of transistor (a component of computers) has a 2% defective rate. The production is considered a random process where each transistor is independent of the others.

(a) What is the probability that the 10th transistor produced is the first with a defect?

```
defective_rate <- .02
prob_ten_fail <- ((1 - defective_rate)^(10 - 1)) * defective_rate
paste("The probability that the 10th transistor produced is the first with a defect is",
      percent(prob_ten_fail))
```

[1] "The probability that the 10th transistor produced is the first with a defect is 1.67%"

(b) What is the probability that the machine produces no defective transistors in a batch of 100?

```
prob_100_no_fail <- ((1 - defective_rate)^(100 - 1)) * (1 - defective_rate)
paste("The probability that the machine produces no defective transistors in a batch of 100 is",
      percent(prob_100_no_fail))
```

```
## [1] "The probability that the machine produces no defective transistors in a batch of 100 is 13.3%"
```

(c) On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?

```
count_to_first_defect <- 1/defective_rate
paste("We would expect on average that",
      count_to_first_defect,
      "would be produced before the first defect would appear",
      sep =" ")
```

[1] "We would expect on average that 50 would be produced before the first defect would appear"

```
standard_deviation <- sqrt((1 - defective_rate)/(defective_rate ^ 2))
paste("The standard deviation is",
      round(standard_deviation,4), sep = " " )
```

[1] "The standard deviation is 49.4975"

(d) Another machine that also produces transistors has a 5% defective rate where each transistor is produced independent of the others. On average how many transistors would you expect to be produced with this machine before the first with a defect? What is the standard deviation?

```
new_defective_rate <- .05
new_count_to_first_defect <- 1/new_defective_rate
paste("We would expect on average that",
      new_count_to_first_defect,
      "would be produced before the first defect would appear",
      sep =" ")
```

[1] "We would expect on average that 20 would be produced before the first defect would appear"

```
new_standard_deviation <- sqrt((1 - new_defective_rate)/(new_defective_rate ^ 2))
paste("The standard deviation is",
      round(new_standard_deviation,4), sep = " " )
```

[1] "The standard deviation is 19.4936"

(e) Based on your answers to parts (c) and (d), how does increasing the probability of an event

affect the mean and standard deviation of the wait time until success?

> By increasing the likelihood of a defect we see that defects will on average appear sooner and more often. The standard deviation is also smaller because the event takes place more often.

3.38 Male children.

While it is often assumed that the probabilities of having a boy or a girl are the same, the actual probability of having a boy is slightly higher at 0.51. Suppose a couple plans to have 3 kids.

(a) Use the binomial model to calculate the probability that two of them will be boys.

```
n <- 3
k <- 2
p <- .51

answer_a <- choose(3,2) * ((p^k)*((1-p)^(n-k)))
answer_a
```

[1] 0.382347

(b) Write out all possible orderings of 3 children, 2 of whom are boys. Use these scenarios to calculate the same probability from part (a) but using the addition rule for disjoint outcomes. Confirm that your answers from parts (a) and (b) match.

```
B <- 'boy'
G <- 'girl'

a <- c(B,B,G)
b <- c(B,G,B)
c <- c(G,B,B)

X <- rbind(a)
X <- rbind(X, b)
X <- rbind(X, c)
X <- matrix(X, ncol = 3)
colnames(X) <- c('child_1', 'child_2', 'child_3')
X
```

```
##      child_1 child_2 child_3
## [1,] "boy"   "boy"   "girl"
## [2,] "boy"   "girl"  "boy"
## [3,] "girl"  "boy"   "boy"
```

```
B <- .51
G <- .49

a <- c(B,B,G)
b <- c(B,G,B)
c <- c(G,B,B)

X <- rbind(a)
X <- rbind(X,b)
X <- rbind(X,c)
X <- matrix(X, nrow = 3)
X <- cbind(X, seq(1:3))
X[1,4] <- X[1,1] * X[1,2] * X[1,3]
X[2,4] <- X[2,1] * X[2,2] * X[2,3]
X[3,4] <- X[3,1] * X[3,2] * X[3,3]
answer_b <- sum(X[,4])
answer_b
```

```
## [1] 0.382347
```

```
paste("The answer to if a is equal to b is:",
      answer_a == answer_b)
```

```
## [1] "The answer to if a is equal to b is: TRUE"
```

(c) If we wanted to calculate the probability that a couple who plans to have 8 kids will have 3 boys, briefly describe why the approach from part (b) would be more tedious than the approach from part (a).

> The approach in b is a very manual process which can be prone to errors. Also, the approach in b does not scale easily for larger problems. In approach (a) the values are implemented in a formula so that the calculations can be performed cleanly and easily without a number of steps allowing for errors.

3.42 Serving in volleyball.

A not-so-skilled volleyball player has a 15% chance of making the serve, which involves hitting the ball so it passes over the net on a trajectory such that it will land in the opposing team's court. Suppose that her serves are independent of each other.

(a) What is the probability that on the 10th try she will make her 3rd successful serve?

```
n <- 10
k <- 3
p <- .15
answer <- choose(9,2) * ((p^k)*((1-p)^(n-k)))
paste("The probability that on the 10th try she will make her 3rd successful serve is",
      percent(answer))
```

[1] "The probability that on the 10th try she will make her 3rd successful serve is 3.9%"

(b) Suppose she has made two successful serves in nine attempts. What is the probability that her 10th serve will be successful?

The probability of a successful serve on the 10th serve is 15% because the probability of a successful serve is independent of the previous serve.

(c) Even though parts (a) and (b) discuss the same scenario, the probabilities you calculated should be different. Can you explain the reason for this discrepancy?

In (a) we are calculating the probability of 3 successful serves with the last successful serve on the 10th serve. Whereas, in (b) we are looking at the probability of a successful serve on the 10th serve when two successful serves have taken place. So its natural that (a) will be different because we are looking at the probability of a sequence taking place that meets the criteria and (b) is really asking for more of what is the probability of a single success.